CLUSTERING BIOLOGICAL ANNOTATIONS AND GENE EXPRESSION DATA TO IDENTIFY PUTATIVELY CO-REGULATED BIOLOGICAL PROCESSES

CORNELIU HENEGAR

1. Inserm, U755 Nutriomique, 75004 Paris, France;
2. Pierre and Marie Curie Paris 6 University, Faculty of Medicine Les Cordeliers, 75004 Paris, France;
3. AP-HP, Hôtel-Dieu Hospital, Nutrition department, 1 Place du parvis Notre-Dame, 75004 Paris, France.
corneliu@henegar.info

RAFFAELLA CANCELLO

1. Inserm, U755 Nutriomique, 75004 Paris, France;
2. Pierre and Marie Curie Paris 6 University, Faculty of Medicine Les Cordeliers, 75004 Paris, France;
3. AP-HP, Hôtel-Dieu Hospital, Nutrition department, 1 Place du parvis Notre-Dame, 75004 Paris, France.
raffaella.cancello@ea3502.org

SOPHIE ROME

4. UMR Inserm U449/INRA U1235, 69008 Lyon, France;
5. Human Nutrition Research Centre, 69008 Lyon, France;
6. Lyon 1 University, Laennec Medical Faculty, 8 rue G. Paradin, 69008 Lyon, France.
srome@univ-lyon1.fr

HUBERT VIDAL

4. UMR Inserm U449/INRA U1235, 69008 Lyon, France;
5. Human Nutrition Research Centre, 69008 Lyon, France;
6. Lyon 1 University, Laennec Medical Faculty, 8 rue G. Paradin, 69008 Lyon, France.
hubert.vidal@laennec.univ-lyon1.fr

KARINE CLÉMENT

1. Inserm, U755 Nutriomique, 75004 Paris, France;
2. Pierre and Marie Curie Paris 6 University, Faculty of Medicine Les Cordeliers, 75004 Paris, France;
3. AP-HP, Hôtel-Dieu Hospital, Nutrition department, 1 Place du parvis Notre-Dame, 75004 Paris, France.
karine.clement@htd.aphp.fr

JEAN-DANIEL ZUCKER

1. Inserm, U755 Nutriomique, 75004 Paris, France;
7. LIM&BIO EA3969, Paris Nord University 74, rue Marcel Cachin, 93017 Bobigny cedex, France.
zucker@smbh.univ-paris13.fr

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Functional profiling is a key step of microarray gene expression data analysis. Identifying co-regulated biological processes could help for better understanding of underlying biological interactions within the studied biological frame. We present herein an original approach designed to search for putatively co-regulated biological processes sharing a significant number of co-expressed

genes. An R language implementation named "FunCluster" was built and tested on two gene expression data sets. A discriminatory functional analysis of the first data set, related to experiments performed on separated adipocytes and stroma vascular fraction cells of human white adipose tissue, highlighted the prevalent role of non adipose cells in the synthesis of inflammatory and immunity molecules in human adiposity. On the second data set, resulting from a model investigating insulin coordinated regulation of gene expression in human skeletal muscle, FunCluster analysis spotlighted novel functional classes of putatively co-regulated biological processes related to protein metabolism and the regulation of muscular contraction. Supplementary information about the FunCluster tool is available on-line at *http://corneliu.henegar.info/FunCluster.htm*

Keywords: gene expression pattern analysis; functional profiling of gene expression; computational biology;

1. Introduction

Microarray technologies are producing important amounts of RNA expression data related to a great variety of biological models, thus promoting the understanding of gene regulation in a variety of conditions. Relying on dedicated statistical approaches, "target genes", corresponding to significantly up or down-regulated RNA transcripts, are identified through differential expression assessment.[1,2] Considered as a major challenge, the translation from RNA expression data to relevant biological mechanisms is indispensable for the comprehension of the underlying biological phenomena.[3,4]

Although text mining techniques designed to extract functional information about genes from scientific text sources proved very useful for automated functional profiling of gene expression data,[5,6] most of the currently available tools are relying on functional annotations stored in curated international genomic databases. Within these resources the biological information is represented in a standardized formal way by using controlled terminologies to associate functional annotations to genes according to the current status of the biological knowledge.

The most widely used terminological system for the annotation of gene and gene products, developed and maintained by the Gene Ontology (GO) Consortium, is structuring functional annotations in a lattice model which reflects the relationship between the biological categories and associated genes.[7] Within GO relevant biological information is partitioned inside three distinct ontologies – Biological Process, Cellular Component and Molecular Function – the lattice structure allowing for the representation of conceptual information with various levels of precision. Within other annotation systems, such as the KEGG Pathway database provided by the Kyoto Encyclopedia for Genes and Genomes (KEGG), the functional assignment aims at linking sets of genes in the genome with a network of interacting molecules in the cell in order to represent high order biological information related to the metabolic and regulatory pathways.[8]

An increasing number of tools, developed during the last years, are proposing different approaches for automatic annotation and functional profiling of microarray data (selected examples: Berriz et al. 2003; Dennis et al. 2003; Zeeberg et al. 2003; Al-Shahrour et al. 2004; Beissbarth et al. 2004; Pandey et al. 2004; Robinson et al. 2004; Smid et al. 2004; Zhang et al. 2004).[4,9-16] Most usually, the functional profiling of gene expression data sets results in a list of functional categories (GO categories, KEGG pathways, etc.) considered

as relevant for the studied biological model. In order to facilitate human analysis some of the aforementioned tools are offering special functionalities like assessing the overrepresentation significance for each functional category,[13,15,16] grouping functional categories based on their position within an ontological hierarchy,[15] analyzing annotations at a predefined level of conceptual precision,[16] or weighting functional categories by the expression intensity of their annotated transcripts.[17]

Besides the functional profiling another important analytical trend is related to the study of genetic regulatory mechanisms which relies on numerous approaches for clustering microarray expression data in search for co-regulated transcripts.[18] In spite of their popularity the usefulness of these approaches is actually limited by the variable granularity of the resulting transcript clusters and the absence of objective criteria for determining cluster boundaries on which depends their biological interpretability. To overcome such limitations integrated approaches, relying on additional biological information (*e.g.* known involvement in some biological regulatory networks), were developed in order to improve the biological relevance of the clustering results.[19]

Acknowledging the advantage of a better integration of expression data with available biological information about transcripts functional roles we present herein an original approach of microarray data analysis designed to provide better insights into the regulatory mechanisms controlling the dynamics of biological processes. In order to test the potential benefit of this approach we developed an algorithm, implemented into an R package named FunCluster, and used it to analyze two in-house gene expression data sets. Based on an adapted clustering procedure of microarray expression data used in conjunction with an automated annotation procedure, FunCluster identifies relevant GO categories and KEGG pathways sharing a significant number of co-expressed RNA transcripts and groups them into functional classes of putatively co-regulated biological processes.

2. System and methods

2.1 *Rationale*

The study of regulatory networks and co-regulated biological processes is a key issue for the understanding of the underlying biological phenomena.[19,20] It is currently accepted that RNA transcripts sharing similar expression patterns throughout a series of different experimental conditions are likely to be regulated via closely related biological mechanisms.[18] Yet there is still no definite way, to our knowledge, to make similar inferences about the existence of some common regulatory mechanisms controlling biological processes or functions. This being so, accumulating recent evidence[20-22] is suggesting the possibility of some regulatory inferences at the level of biological processes and functions based on the expression similarity of the genes controlling these processes. Among others, Alocco *et al.*[20] showed that RNA transcripts with high levels of expression similarity, corresponding to strongly correlated expression patterns (*e.g.* with

a correlation coefficient > 0.84), are more likely to have their promoter regions bound by common transcription factors and, concomitantly, to share similar functional roles.
Starting from these premises we designed an original approach of processing microarray data, complementary to the available functional profiling tools, aiming to identify functional classes of putatively co-regulated biological processes, functions or metabolic pathways and to rigorously assess their significance for the studied biological model.

2.2 *System overview*

FunCluster's functional analysis relies on a pre-selection of "target" genes considered as relevant for the studied biological frame on the basis of a significant variation of their expression level. As FunCluster does not assess itself the significance of gene expression variation this has to be done prior to the functional analysis by using one of the tools currently available for performing this task under various settings and conditions, as the SAM[1] or VarMixt[2] algorithms or the approaches implemented into the Bioconductor packages (available at *http://www.bioconductor.org/*).
Although FunCluster was specifically designed for the analysis of data resulting from dichotomous microarray models, in which expression measurements are performed in a number of subjects in relation to a punctual biological intervention, this approach can be used in relation with a variety of other experimental settings. Two such possibilities are illustrated in this paper. Starting from one or two lists of differentially expressed transcripts FunCluster's functional analysis follows three main steps among which the first two are independent while the third one is relying on the two previous ones:

(1) partitioning expression data into clusters of strongly co-expressed genes by using predefined quality criteria for significant correlation of their expression profiles;
(2) automated annotation of relevant genes and overrepresentation analysis (*e.g.* gene enrichment) of biological categories annotating the analyzed data set;
(3) identifying putatively co-regulated biological processes and functions, sharing a significant number of strongly co-expressed genes, and grouping them into relevant functional classes.

2.3 *Partitioning expression data into clusters of strongly co-expressed genes*

Often used as a basis for further computational analysis, clustering gene expression data is a common exploratory procedure of gene expression patterns in microarray data sets. Based on several types of similarity measures for gene expression profiles and on a number of predefined constraints clustering procedures are providing partitions of disjoint groups of co-expressed genes (clusters). Among various similarity measures the standard correlation coefficient between series of expression measurements is known to capture the similarity between the "shapes" of expression profiles without paying attention to the magnitude of gene expression. As a consequence it has been proposed as a good measure of the intuitive biological notion of what it means for two genes to be co-expressed.[18] Because of its lack of assumptions on data distribution, resulting in a broad applicability on various data sets, we have chosen the pairwise Spearman's Rank

correlation coefficient (R_s) to be FunCluster's default metric for gene expression similarity.

Considering two genes *X* and *Y*, the Spearman's R_s is computed between two vectors of log intensity ratios $log_2(Cy5/Cy3)$ measured across a set of *n* different conditions, after pairwise elimination of missing values, as:

$$Rs(X,Y) = 1 - \frac{6\left(\sum_{i=1}^{n} di^2\right)}{n(n^2-1)} \qquad (1)$$

In relation (1) d_i is the difference between the ranks of the two variables *x* and *y* calculated as:

$$di = Rank(xi) - Rank(yi) \qquad (2)$$

The statistical significance of the correlation is assessed by checking all rank permutations within each of the two series of values and then counting the fraction for which the Spearman's Rank correlation coefficient R_s' is more extreme than the R_s initially found.

One major issue concerning clustering techniques is the establishment of discriminative cluster boundaries without making a priori assumptions about the number of relevant clusters existing in a data set.[19] For this purpose classical cluster validity assessment techniques are using quality measures based on external or internal criteria in order to evaluate the compactness and the separation features of a resulting clustering scheme.[23] Other techniques are designed around relative criteria which rely on unsupervised learning techniques evaluating clusters stability within an ensemble of partitions.[24] More specific approaches, aiming to improve the biological interpretability of clustering results, are using supplementary information about genes in order to promote biologically relevant clustering. Most usually such information relates to the sharing of similar biological roles[17] or a common involvement in known biological regulatory networks.[19,25,26] In our case, in order to support the type of regulatory assertions targeted by FunCluster, a specific quality criterion for clustering biologically significant co-expressed genes was derived based on strong evidence from pre-cited recent analysis[20]. This criterion combines a threshold for the Spearman's correlation coefficient between expression patterns ($R_s \geq 0.85$), with a FDR adjusted p-value of statistical significance ($p \leq 0.05$).

Under these circumstances FunCluster builds clusters of strongly correlated genes from available expression data through a greedy heuristic approach. A preliminary step before initiating the clustering procedure is the ranking of relevant genes by the decreasing order of the statistical significance of their differential expression. Besides the fact that it guarantees the reproducibility of clustering results this step allows taking into account the magnitude of genes differential expression. Thus, starting with the first gene (*e.g.* the one with the most significant differential expression), considered as a *clustering seed*, FunCluster computes all the correlation coefficients R_s between its expression profile and

those of the other genes available in the data matrix. Afterwards, a first gene cluster is built by grouping together all the genes which satisfy the aforementioned quality criterion in relation to the considered seed. If no gene satisfies the clustering criterion than the reference seed is eliminated from the list of genes and the clustering process is reiterated. Once a cluster is created newly clustered genes are eliminated from the data matrix and the operation is repeated until no gene is left unclustered. As repeated testing for correlation significance multiplies the risk of false positive results FunCluster use various methods for p-values correction as described in section 2.6. Expression data partitioning results eventually in a list of strongly co-expressed gene clusters which are required by the last step of the functional analysis.

2.4 *Automated annotation of expression data and overrepresentation analysis of biological categories*

During a second step, independent from the first one, FunCluster uses an automated annotation procedure to identify biological processes, functions or metabolic pathways which are relevant for the studied biological frame. Two types of genomic annotations are currently available for FunCluster's functional profiling: GO categories belonging to Biological Process, Cellular Component and Molecular Function ontologies and KEGG metabolic and regulatory pathways. The automated annotation procedure relies on organism specific genomic annotations which are retrieved from their original web sources through an automated update routine and stored locally for further processing.

A preliminary step required by the automated annotation procedure is the conversion of gene identifiers (*e.g.* most usually cDNA IDs) to a standardized gene accession (SGA) system. This allows avoiding well known redundancies of cDNA probes in microarray data and assures a correct overrepresentation analysis of the biological categories annotating analyzed genes. As default SGA system FunCluster is using EntrezGene GeneID identifiers (available at *http://www.ncbi.nlm.nih.gov/entrez*).[27]

Afterwards an automated annotation procedure is performed by retrieving from the aforementioned resources all biological categories annotating relevant genes. In order to support a more precise functional analysis the automated annotation procedure favors the selection of the most specific GO categories against more general ones by restricting annotations inheritance within the GO ontological lattice. Two approaches are thus provided for relating genes to GO categories. The most limitative one restricts the automated annotation procedure to those genes directly annotated by each biological category (*e.g.* its direct instances). While this approach reduces biological noise to a minimum its restrictive behavior may result sometimes in filtering too much meaningful biological information. Therefore a second approach was designed by considering for each GO category, besides its directly annotated genes, those genes which are directly annotated by one of its directly subsumed categories within the GO ontological lattice.

The automated annotation procedure is followed by an overrepresentation analysis which aims to identify context relevant biological categories through a separate analysis of each of the three GO ontologies as well as of the KEGG metabolic and regulatory pathways. Thus for each available biological category a 2 x 2 matrix of frequencies (Table 1) is

established by successively counting the number of its occurrences within the analyzed list of differentially expressed genes (*a* in Table 1), as well as within a reference list of genes including all the genes available before the selection of differentially expressed ones ($C_1 = a + c$). For calculation purposes FunCluster computes also the total number of differentially expressed genes ($R_1 = a + b$) and the number of genes belonging to the reference list ($N = R_1 + R_2 = C_1 + C_2$) for which at least one annotation was available.

Table 1. The contingency table established for an annotation *A* regrouping its observed occurrences within the analyzed list of genes *L* and within the reference list of genes *N*.

	A+	*A-*	
L+	*a*	*b*	R_1
L-	*c*	*d*	R_2
	C_1	C_2	*N*

As previously suggested,[3,13,16] significantly overrepresented biological categories are identified by assessing the statistical relevance of gene enrichment for each GO or KEGG category. Thus FunCluster computes a significance p-value by applying a one-sided Fisher exact test to the matrix of frequencies. The test statistics relies on the hypergeometric distribution which allows calculating the probability to observe the actual matrix of frequencies among all possible combinations, as:

$$P_{cutoff} = \frac{R1!R2!C1!C2!}{a!b!c!d!N!} \qquad (3)$$

In relation (3) R_1, R_2 are the sums of the frequencies by row, C_1, C_2 the sums of the frequencies by column, *N* the total sum of the frequency table and *a*, *b*, *c*, *d* the elements of the observed matrix of frequencies. The significance p-value of gene enrichment is computed as the sum of cutoff probabilities for all the theoretic matrices corresponding to a higher enrichment than the observed one. In the end the overrepresentation analysis results in a list of significantly enriched functional categories (considered as relevant for the studied biological frame) ranked in the order of their statistical significance.

2.5 *Co-clustering relevant biological annotations and gene expression data*

During the third analytical step FunCluster groups together relevant biological categories which share a significant number of strongly co-expressed genes through an agglomerative procedure relying on the partition of genes into clusters of strongly correlated expression profiles (step 1). This procedure results into a list of functional classes grouping putatively co-regulated biological processes, thus illustrating context specific regulatory patterns not only at the gene level but also at the higher conceptual level of biological functions and processes.

The agglomerative clustering procedure starts with the first item of an ordered list of relevant biological categories (step 2), thus considered as a clustering seed, and tests it for significant association against each of the other categories in the list. The association between the considered clustering seed and each of the other biological categories is assessed by identifying and counting all co-expressed genes shared by the two categories through a confrontation procedure against each of the co-expressed gene clusters previously found (step 1). Afterwards a significance p-value of the association between the two categories is computed by using a one-sided Fisher exact test which evaluates the theoretical probability to see a similar or a bigger number of co-expressed genes shared by the two tested categories. If significant associations are detected a functional cluster is created by grouping the clustering seed with the most significantly associated category. Then the resulting cluster is added to the list of relevant functional categories in replacement of those which were clustered together and the agglomerative procedure is reiterated by considering as seed the newly created cluster. Eventually, when no more significant associations with the new cluster are detected, the functional cluster is removed from the list of unclustered categories and added to the final list of functional classes. Subsequently the clustering procedure is reinitiated by considering as clustering seed the most relevant biological category among the remaining unclustered ones. When all biological categories have been tested for significant associations the clustering process ends and the final list of functional classes is saved in local files for further human interpretation.

2.6 *Handling false positive results*

As repeated testing for statistical significance multiplies the risk of false positive results (type I errors) FunCluster provides several methods for adjusting computed p-values either by controlling the family wise error rate (FWER) with the Hommel (1988) correction, or by using one of the methods available for the false discovery rate (FDR) control. By default p-values adjustment is computed with the Benjamini and Hochberg (2001) method for FDR control which is known to provide a good balance between statistical power and computational cost.[28] The other methods implemented in FunCluster for controlling the FDR are the Benjamini and Yekutieli (2001) method,[29] more conservative but offering a stronger control under arbitrary dependency of test statistics, and the Storey (2002) method which provides increased accuracy and power over the Benjamini and Hochberg approach while being more computationally expensive.[30]

3. Results

FunCluster can analyze various types of microarray data sets resulting from experimental models involving either series of successive distinct biological conditions or series of different individuals subjected to the same experimental condition. Besides the functional profiling of a single list of significant genes FunCluster can equally perform a discriminatory functional analysis of two lists of genes. Two examples, one for each such possibility, will be given in this section.

Among the parameters* which can be used to tune-up FunCluster's analytical procedure the most important one is the threshold of the Spearman's R_s on which depends the clustering of gene expression profiles (step 1). Therefore we assessed FunCluster's greedy heuristic (described in subsection 2.3) by comparing it to a classical approach combining a hierarchical agglomerative clustering approach with Silhouette computations in order to identify an optimal partition of gene clusters.[31] Table 2 illustrates the average intracluster Spearman's R_s yielded by these two clustering approaches applied on the two microarray data sets presented in subsection 3.2. Besides the computation costs of the combined classical approach, which may increase rapidly depending on the number of elements to be clustered, the resulting average intracluster R_s show important variations, in some cases well below the significance threshold required by the targeted regulatory inferences ($R_s \geq 0.85$). By contrast the greedy heuristic yielded a more stable intracluster R_s constantly situated within the targeted range of significance required by FunCluster's regulatory inferences.

Table 2. The average intracluster Spearman's R_s yielded by FunCluster's greedy heuristic compared to a classical approach combining hierarchical agglomerative clustering with Silhouette computation

RNA transcript data sets	FunCluster's greedy heuristic R_s (mean ± SD* [number of clusters])		Combined agglomerative approach R_s (mean ± SD* [number of clusters]) | {Silhouette}**
	Threshold $R_s \geq 0.85$	Threshold $R_s \geq 0.95$	
Insulin up-regulated	0.95 ± 0.07 [18]	0.99 ± 0.03 [35]	0.77 ± 0.23 [6] | {0.43}
Insulin down-regulated	0.92 ± 0.08 [7]	0.99 ± 0.01 [14]	0.78 ± 0.21 [2] | {0.71}
Adipocytes	0.88 ± 0.11 [57]	0.96 ± 0.07 [140]	0.58 ± 0.31 [3] | {0.54}
SVF	0.88 ± 0.10 [57]	0.98 ± 0.04 [214]	0.44 ± 0.38 [2] | {0.43}

* SD – standard deviation of the mean intracluster Spearman's R_s; ** Silhouette index of the optimal partition of clusters

3.2 *Results in application*

In order to evaluate the potential benefit of FunCluster's automated functional analysis we tested it on two distinct gene expression data sets.† The first one resulted from microarray experiments performed on human white adipose tissue after the separation of its two cellular components: mature adipocytes and stroma vascular fraction (SVF) cells, the "non adipose" component of the tissue. The purpose of this experimental model was to distinguish the two cellular fractions from a functional perspective and, in the meantime, to help establish the contribution of "adipose" and "non adipose" cells in the expression of inflammatory molecules in morbid obesity.[32] Therefore we performed a discriminatory functional analysis of two lists of genes identified to be specifically expressed within adipocytes and SVF cells respectively.

* More details about the format of the data files and the available settings are provided with the FunCluster package which can be downloaded either from FunCluster's webpage *http://corneliu.henegar.info/FunCluster.htm* or from CRAN repositories at *http://www.r-project.org/*.

† Detailed results of the functional analysis of these two data sets are provided as supplementary data from FunCluster's webpage.

FunCluster analysis revealed that a majority of the biological categories which characterize genes expressed in adipose cells were related to "energy metabolism" and "lipid and glucose metabolism", all known metabolic processes specific to mature adipocytes (Fig. 1). Thus the functional pattern of human adipocytes depicted by our experiments is in agreement with previously published gene profiling studies in humans[33] and in small mammals adipose cells.[34]

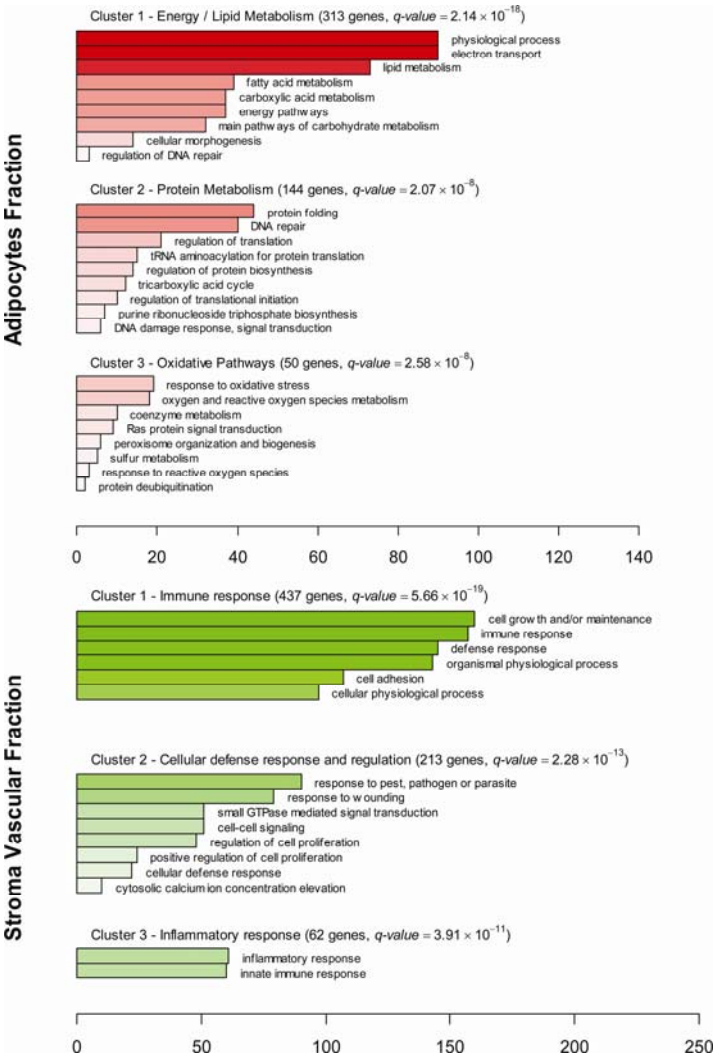

Fig. 1. The most significant functional clusters of GO Biological Process categories yielded by the discriminatory analysis of differentially expressed genes in mature adipocytes and stroma vascular fraction cells extracted from white adipose tissue of morbidly obese human subjects.

By contrast the most significant biological categories which characterize SVF specifically expressed genes were related to "inflammatory" or "immunity" processes (Fig. 1).

Indeed, while it has been convincingly shown that mature adipose cells are also capable to express and secrete inflammatory molecules,[35,36] our analysis clearly underlined the preponderant role of SVF cells in immunity and inflammation-related processes characterizing human obesity.[37,38]

The second data set resulted from a previous study investigating insulin coordinated regulation of gene expression in human skeletal muscle during a 3-hour hyperinsulinemic euglycemic clamp.[39] Seven main functional groups were originally identified (Table 3) based on a tedious classification of genes relying on manual queries of available databases (SOURCE, OMIM, PubMed).

Table 3. A comparative presentation of the main classes of biological processes up-regulated by insulin in the human skeletal muscle reported by *Rome et al, 2003* versus those yielded by FunCluster's functional analysis

Main functional classes	Gene space covering (%)		Related FunCluster functional clusters	
	Rome et al, 2003	FunCluster*	Clusters ranks	Significance range**
Transcriptional and translational regulation	29	22	1, 3, 4, 5, 10, 12, 15	$1.41\ 10^{-9} – 2.61\ 10^{-2}$
Muscular contraction	-	16	1, 4, 6, 9	$1.41\ 10^{-9} – 4.33\ 10^{-4}$
Macromolecule and proteic biosynthesis	-	23	1, 12	$1.41\ 10^{-9} – 5.83\ 10^{-3}$
Ubiquitin-proteasome pathway	7	16	2, 7	$2.79\ 10^{-8} – 2.99\ 10^{-4}$
Intermediary and energy metabolism	14	23	3, 4, 6, 17, 18	$1.94\ 10^{-7} – 3.24\ 10^{-2}$
Cytoskeleton and vesicle traffic	9	9	4, 6, 9	$5.90\ 10^{-6} – 4.34\ 10^{-4}$
Intracellular signaling	12	10	7, 9, 13, 18, 19	$2.99\ 10^{-4} – 3.24\ 10^{-2}$
Receptors carriers and transporters	8	10	8, 9, 11, 12, 13, 14, 17, 18	$2.99\ 10^{-4} – 3.24\ 10^{-2}$
Immune response components	5.5	3	8, 11, 14	$2.99\ 10^{-4} – 1.45\ 10^{-2}$

* Estimation relying on genes annotated by significantly overrepresented GO Biological Process categories; ** *Q-values* computed with the Benjamini & Hochberg (2001) FDR method estimating the significance the gene space coverage of individual clusters.

Besides confirming some of the original findings FunCluster analysis identified novel functional classes, which were missed by the manual classification, and quantified their relevance for the studied experimental model in a rigorous manner. Indeed, automated analysis yielded significantly overrepresented biological categories related to protein biosynthesis and metabolism (Cluster 1 in Fig. 2) as well as to the regulation of skeletal muscle contraction (Cluster 3 in Fig. 2).

One of the most important biological effects of the insulin resides in a strong stimulation of protein synthesis.[40] FunCluster's analysis not only underlined the impact of the hormone on protein metabolism but it also spotlighted the existence of insulin dependent regulatory mechanisms controlling key processes in human muscle (Fig. 2). Thus cluster

1 in Fig. 2 illustrates the existence of common regulatory mechanisms involved in transcription and translation processes, protein synthesis and myogenesis.

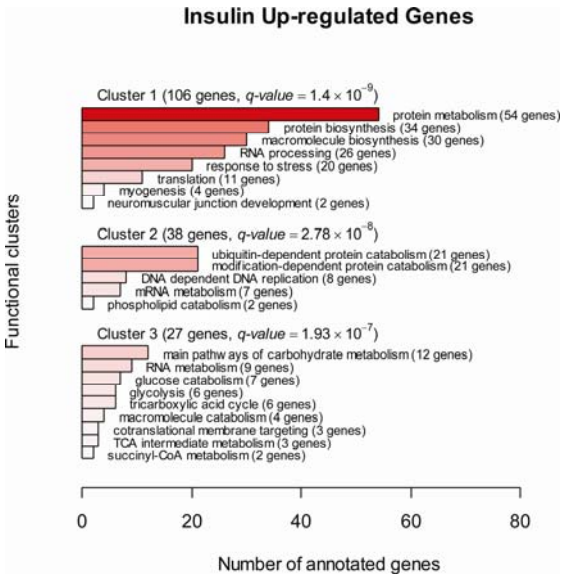

Fig. 2. The three most significant functional clusters of GO Biological Process categories resulting from FunCluster functional analysis of insulin up-regulated genes in human muscle.

Insulin impact on muscular contraction is supported by a large amount of recent data describing the relationship between insulin action, physical exercise and muscle contraction in vivo. Both contraction and insulin increase glucose uptake and oxidation and affect lipid metabolism in muscle.[41,42] FunCluster confirmed that an important component of insulin action in the skeletal muscle is the ubiquitin-proteasome pathway (Cluster 2 in Fig. 2) which is the major pathway of selective protein degradation in eukaryotic cells.[43] Again, FunCluster pointed out novel insulin co-regulated biological processes related to the proteasome-ubiquitin pathway and the regulation of transcription and RNA metabolism which formed a cluster including 38 genes (Cluster 2 in Fig. 2). Recently, this pathway has emerged as a new important regulator of the insulin signaling mechanism.[44,45]

Finally the role of insulin in the intermediary and energy metabolism of the skeletal muscle has been confirmed (Cluster 3 in Fig. 2) in agreement with previous evidence showing a hormone dependent increase in the uptake of glucose and its storage into glycogen.[46]

3.3 *Comparison with available functional profiling tools*

In order to better illustrate the specificity of FunCluster's functional inferences we realized a comparative applicative assessment of three other tools which rely on genomic annotations for functional profiling of microarray expression data. Among them the

FatiGO & FatiWise group of tools[16] was chosen to exemplify the automated functional profiling of gene expression data based on GO (FatiGO) or KEGG (FatiWise) genomic annotations. On its turn the GOstat tool[15] illustrates the use of the conceptual information stored within the GO ontological lattice for grouping closely related and relevant biological categories. Detailed comparative results obtained with these tools on the two previously described data sets are available as supplementary data on the FunCluster's website.

As it can be seen from these data the FatiGO functional analysis resulted in some unbalanced and heterogeneous biological information in spite of relying exclusively on GO categories belonging to the same level of the ontological structure. For instance, while the discriminatory analysis of the adipose tissue data set identified metabolic processes (represented by the GO category "metabolism") as being more characteristic of mature adipocytes, the SVF cells were poorly characterized by GO categories as "response to stimulus" and "cell communication". On the same data set the FatiWise tool yielded a more homogeneous characterization of the functional profile of the two cellular fractions, in accord with FunCluster's results. Thus relevant adipocyte specific pathways as "oxidative phosphorylation", "fatty acid metabolism" or "pyruvate metabolism" were distinguished from SVF biological themes as "complement and coagulation cascades" or "apoptosis", although without providing any insights on the more profound biological regulatory mechanisms controlling these processes.

By contrast the GoStat tool allowed for a more detailed description of the biological themes overrepresented in the two analyzed data sets, most probably because of the absence of constraints related to the ontological level of the analyzed GO categories. For example GO categories as "generation of precursor metabolites and energy", "electron transport", "coenzyme catabolism", "energy derivation by oxidation of organic compounds", "metabolism", "tricarboxylic acid cycle" were recognized to be specific of mature adipocytes while categories as "immune response", "defense response", "cellular defense response", "chemotaxis", "humoral defense mechanism" characterized SVF cells. Nevertheless, in spite of its obvious benefits, the grouping of relevant GO categories based on their ontological proximity lacks a genuine context specific biological relevance.

4. Discussion and perspectives

FunCluster's functional profiling of gene expression data belongs to the latest trend of integrated approaches aiming to reunite a maximum of significant information in order to support a finer, more systematic and more relevant biological analysis of microarray data. Like other automated approaches it provides all the advantages of an automated annotation combined with a rigorous statistical assessment of relevant biological categories overrepresented within gene expression data sets. However, the original approach of functional profiling described here distinguishes itself from other currently available knowledge guided analytical approaches in a number of key aspects.

The main conceptual innovation of FunCluster's approach is illustrated by the design of its combined clustering technique which tries to complement the functional information about genes biological roles by relating it to the expression data in order to enrich its biological relevance. The results presented in this paper appear to confirm the advantages of this approach which allowed a more precise functional analysis spotlighting, besides main biological themes, some of the regulatory mechanisms controlling biological functions and processes.

Also, by contrast to other approaches which are using the conceptual information stored within the GO ontological structure as an exclusive base for regrouping relevant biological categories FunCluster makes only a limited use of the local structure of GO. This use is reserved for the purpose of gene enrichment computations which are restricted, as previously explained, to directly subsumed concepts within the ontological network. While this restriction resulted in an increased biological precision of the functional analysis it could not avoid completely the theme redundancy resulting from the sometimes problematic conceptual transitivity of the GO ontological lattice.

Another potential utility of FunCluster's analytical approach would be to consider it as a starting point for novel "guilt by association" functional inferences,[21,47] which could presumably allow to enrich functional information on genes and expressed sequence tags with yet unknown biological role. Future refinements of the functional clustering procedure could be assumed by integrating an analysis of genes regulatory regions relying on common recognition motifs in order to strengthen the relevance of the co-regulation inferences based upon gene expression data.

Acknowledgments

This work was supported by the Institut National de la Santé et de la Recherche Médicale (INSERM), and the Assistance Publique – Hôpitaux de Paris. The authors received a grant from Paris VI University (BQR, bonus quality research), and from INSERM (PRNH, Research Program on Human Nutrition). Corneliu Henegar is funded by INSERM, ADR Paris VI, Saint-Antoine, Paris, and he also received a grant from the Assistance Publique – Hôpitaux de Paris.

References

1. V. G. Tusher, R. Tibshirani and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc Natl Acad Sci U S A* **98**(9), 5116-21 (2001).
2. P. Delmar, S. Robin and J. J. Daudin, "VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data," *Bioinformatics* **21**(4), 502-8 (2005).
3. S. Draghici, P. Khatri, R. P. Martins, G. C. Ostermeier and S. A. Krawetz, "Global functional profiling of gene expression," *Genomics* **81**(2), 98-104 (2003).
4. P. N. Robinson, A. Wollstein, U. Bohme and B. Beattie, "Ontologizing gene-expression microarray data: characterizing clusters with Gene Ontology," *Bioinformatics* **20**(6), 979-81 (2004).
5. D. Chaussabel, A. Sher, "Mining microarray expression data by literature profiling," *Genome Biol* **3**(10), RESEARCH0055 (2002).
6. S. Raychaudhuri, R. B. Altman, "A literature-based method for assessing the functional coherence of a gene group," *Bioinformatics* **19**(3), 396-401 (2003).

7. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet* **25**(1), 25-9 (2000).
8. M. Kanehisa, S. Goto, S. Kawashima and A. Nakaya, "The KEGG databases at GenomeNet," *Nucleic Acids Res* **30**(1), 42-6 (2002).
9. M. Smid, L. C. Dorssers, "GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms," *Bioinformatics* **20**(16), 2618-25 (2004).
10. G. F. Berriz, O. D. King, B. Bryant, C. Sander and F. P. Roth, "Characterizing gene sets with FuncAssociate," *Bioinformatics* **19**(18), 2502-4 (2003).
11. B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett and J. N. Weinstein, "GoMiner: a resource for biological interpretation of genomic and proteomic data," *Genome Biol* **4**(4), R28 (2003).
12. B. Zhang, D. Schmoyer, S. Kirov and J. Snoddy, "GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies," *BMC Bioinformatics* **5**(1), 16 (2004).
13. G. Dennis Jr, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane and R. A. Lempicki, "DAVID: Database for Annotation, Visualization, and Integrated Discovery," *Genome Biol* **4**(5), P3 (2003).
14. R. Pandey, R. K. Guru and D. W. Mount, "Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data," *Bioinformatics* **20**(13), 2156-8 (2004).
15. T. Beissbarth, T. P. Speed, "GOstat: find statistically overrepresented Gene Ontologies within a group of genes," *Bioinformatics* **20**(9), 1464-5 (2004).
16. F. Al-Shahrour, R. Diaz-Uriarte and J. Dopazo, "FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes," *Bioinformatics* **20**(4), 578-80 (2004).
17. B. Adryan, R. Schuh, "Gene-Ontology-based clustering of gene expression data," *Bioinformatics* **20**(16), 2851-2 (2004).
18. M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc Natl Acad Sci U S A* **95**(25), 14863-8 (1998).
19. D. Hanisch, A. Zien, R. Zimmer and T. Lengauer, "Co-clustering of biological networks and gene expression data," *Bioinformatics* **18 Suppl 1**, S145-54 (2002).
20. D. J. Allocco, I. S. Kohane and A. J. Butte, "Quantifying the relationship between co-expression, co-regulation and gene function," *BMC Bioinformatics* **5**(1), 18 (2004).
21. C. J. Wolfe, I. S. Kohane and A. J. Butte, "Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks," *BMC Bioinformatics* **6**, 227 (2005).
22. P. J. Park, A. J. Butte and I. S. Kohane, "Comparing expression profiles of genes with similar promoter regions," *Bioinformatics* **18**(12), 1576-84 (2002).
23. M. Halkidi, Y. Batistakis and M. Vazirgiannis, "On Clustering Validation Techniques.," *J. Intell. Inf. Syst.* **17**(2-3), 107-45 (2001).
24. A. F. Famili, G. Liu and Z. Liu, "Evaluation and optimization of clustering in gene expression data analysis," *Bioinformatics* **20**(10), 1535-45 (2004).
25. A. Zien, R. Kuffner, R. Zimmer and T. Lengauer, "Analysis of gene expression data with pathway scores," *Proc Int Conf Intell Syst Mol Biol* **8**, 407-17 (2000).
26. D. Hanisch, F. Sohler and R. Zimmer, "ToPNet--an application for interactive analysis of expression data and biological networks," *Bioinformatics* **20**(9), 1470-1 (2004).
27. D. Maglott, J. Ostell, K. D. Pruitt and T. Tatusova, "Entrez Gene: gene-centered information at NCBI," *Nucleic Acids Res* **33**(Database issue), D54-8 (2005).
28. S. Dudoit, J. P. Shaffer and J. C. Boldrick, "Multiple hypothesis testing in microarray

experiments.," *Statistical Science* **18**(1), 71-103 (2003).
29. A. Reiner, D. Yekutieli and Y. Benjamini, "Identifying differentially expressed genes using false discovery rate controlling procedures," *Bioinformatics* **19**(3), 368-75 (2003).
30. J. D. Storey, R. Tibshirani, "Statistical significance for genomewide studies," *Proc Natl Acad Sci U S A* **100**(16), 9440-5 (2003).
31. N. Bolshakova, F. Azuaje, "Cluster validation techniques for genome expression data," *Signal Processing* **83**, 825-33 (2003).
32. R. Cancello, C. Henegar, N. Viguerie, S. Taleb, C. Poitou, C. Rouault, M. Coupaye, V. Pelloux, D. Hugol, J. L. Bouillot, A. Bouloumie, G. Barbatelli, S. Cinti, P. A. Svensson, G. S. Barsh, J. D. Zucker, A. Basdevant, D. Langin and K. Clement, "Reduction of macrophage infiltration and chemoattractant gene expression changes in white adipose tissue of morbidly obese subjects after surgery-induced weight loss," *Diabetes* **54**(8), 2277-86 (2005).
33. S. Urs, C. Smith, B. Campbell, A. M. Saxton, J. Taylor, B. Zhang, J. Snoddy, B. Jones Voy and N. Moustaid-Moussa, "Gene expression profiling in human preadipocytes and adipocytes by microarray analysis," *J Nutr* **134**(4), 762-70 (2004).
34. J. Gomez-Ambrosi, V. Catalan, A. Diez-Caballero, L. A. Martinez-Cruz, M. J. Gil, J. Garcia-Foncillas, J. A. Cienfuegos, J. Salvador, J. M. Mato and G. Fruhbeck, "Gene expression profile of omental adipose tissue in human obesity," *FASEB J* **18**(1), 215-7 (2004).
35. C. E. Juge-Aubry, E. Somm, R. Chicheportiche, D. Burger, A. Pernin, B. Cuenod-Pittet, P. Quinodoz, V. Giusti, J. M. Dayer and C. A. Meier, "Regulatory effects of interleukin (IL)-1, interferon-beta, and IL-4 on the production of IL-1 receptor antagonist by human adipose tissue," *J Clin Endocrinol Metab* **89**(6), 2652-8 (2004).
36. C. Chiellini, M. Costa, S. E. Novelli, E. Z. Amri, L. Benzi, A. Bertacca, P. Cohen, S. Del Prato, J. M. Friedman and M. Maffei, "Identification of cathepsin K as a novel marker of adiposity in white adipose tissue," *J Cell Physiol* **195**(2), 309-21 (2003).
37. J. N. Fain, S. W. Bahouth and A. K. Madan, "TNFalpha release by the nonfat cells of human adipose tissue," *Int J Obes Relat Metab Disord* **28**(4), 616-22 (2004).
38. J. N. Fain, A. K. Madan, M. L. Hiler, P. Cheema and S. W. Bahouth, "Comparison of the release of adipokines by adipose tissue, adipose tissue matrix, and adipocytes from visceral and subcutaneous abdominal adipose tissues of obese humans," *Endocrinology* **145**(5), 2273-82 (2004).
39. S. Rome, K. Clement, R. Rabasa-Lhoret, E. Loizon, C. Poitou, G. S. Barsh, J. P. Riou, M. Laville and H. Vidal, "Microarray profiling of human skeletal muscle reveals that insulin regulates approximately 800 genes during a hyperinsulinemic clamp," *J Biol Chem* **278**(20), 18063-8 (2003).
40. R. M. O'Brien, D. K. Granner, "Regulation of gene expression by insulin," *Physiol Rev* **76**(4), 1109-61 (1996).
41. F. Tremblay, M. J. Dubois and A. Marette, "Regulation of GLUT4 traffic and function by insulin and contraction in skeletal muscle," *Front Biosci* **8**, d1072-84 (2003).
42. E. A. Richter, J. N. Nielsen, S. B. Jorgensen, C. Frosig and J. F. Wojtaszewski, "Signalling to glucose transport in skeletal muscle during exercise," *Acta Physiol Scand* **178**(4), 329-35 (2003).
43. M. H. Glickman, A. Ciechanover , "The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction," *Physiol Rev* **82**(2), 373-428 (2002).
44. S. Rome, E. Meugnier and H. Vidal, "The ubiquitin-proteasome pathway is a new partner for the control of insulin signaling," *Curr Opin Clin Nutr Metab Care* **7**(3), 249-54 (2004).
45. X. J. Sun, J. L. Goldberg, L. Y. Qiao and J. J. Mitchell, "Insulin-induced insulin receptor substrate-1 degradation is mediated by the proteasome degradation pathway," *Diabetes* **48**(7), 1359-64 (1999).
46. D. E. Kelley, J. P. Reilly, T. Veneman and L. J. Mandarino, "Effects of insulin on skeletal muscle glucose storage, oxidation, and glycolysis in humans," *Am J Physiol* **258**(6 Pt 1), E923-9 (1990).
47. M. D. Robinson, J. Grigull, N. Mohammad and T. R. Hughes, "FunSpec: a web-based cluster interpreter for yeast," *BMC Bioinformatics* **3**(1), 35 (2002).

Corneliu Henegar graduated in Medicine from the Faculty of Medicine at the University of Timisoara, Romania, in 1996. Between 1996 and 1999 he did his residency training in medicine at the University Hospital of Cluj-Napoca, Romania. He also graduated in Informatics in 1998 in Cluj-Napoca, Romania. After coming to France in 1999 he continued his residency training in Internal Medicine at Saint-Antoine Faculty of Medicine of Paris VI University from which he graduated in 2006. In 2003, after one year of research fellowship, he got his Master Degree in Medical Informatics from Paris VI University. Since the fall of 2003 he started a PhD fellowship in Bioinformatics under the guidance of Prof Jean-Daniel Zucker at Paris XIII University. In 2005 he joined the INSERM U755 unit as a fellow researcher. His current scientific work focuses on developing knowledge engineering techniques for data processing in biology and medicine, with a special interest in computational biology and functional analysis of genomic data. He is also interested in philosophical aspects of functional inferences in biology, as he is currently studying to obtain his Master Degree in Philosophy of Science at Paris I "La Sorbonne" University.

Raffaella Cancello was graduated PhD from the Faculty of Medicine at the Polytechnic University of Ancona, Italy, in January 2000. She is specialized in white adipose tissue morphology and cellular biology/physiology of adipocytes. She's presently an INSERM post-doctorant fellow of the Nutrition Department Laboratory, INSERM U755 "Nutriomics", Hôtel-Dieu hospital, Paris, France, directed by the Prof Karine Clément. During her postdoctoral stay she acquired new competencies in microarrays techniques, with the goal of approaching the physiopathology of complex diseases, like human obesity, using gene profiling approach. Her main research interests are the regulation of adiposity signals in white adipose tissue during energy balance variations.

Sophie Rome is a youg scientist at the INRA institut (www.inra.fr). She received her PhD both in molecular biology and bioinformatics in 1996, from the french university Claude Bernard Lyon I. After a postdoctoral stay at the Western University of Australia, and then at the Institut for Molecular Biology of Singapore (IMCB), she was recruited at the french INRA Institut in 1997 to work on the regulation of instestinal amino acid transporters by high protein diets and to develop a new laboratory of molecular biology. In 2001 she joined the Hubert Vidal team at Lyon, France. She is particularly interested in the regulation of gene expression, and focus on how the information that specifies when and where genes are expressed is encoded in genome

sequences and on the role that regulated gene expression plays in human desease development like obesity and type 2 diabetes. She has developped a DNA microarrays platform in Lyon, France, and is particularly interested in software development for microarray data analysis.

Hubert Vidal is Research Director at INSERM and head of a research team in the Unit 449 entitlled "Molecular Mechanisms of Diabetes" in Lyon, France. He presently works on the mechanisms of insulin-resistance during type 2 diabetes and obesity, with particular interests for the regulation of gene expression by nutrients and hormones in vivo in human. Using novel methodologies based on large scale analysis of gene expression using microarrays, his team recently demonstrates the presence of alterations in the regulation of a number of key genes in the skeletal muscle and the adipose tissue of patients with type 2 diabetes. He is author or co-author of 120 publications and was recipient of the Young Researcher Award of the Morgagni prize in 2003 and of the Appolinaire Bouchardat Prize in 2004.

Karine Clément is professor of medicine in the Nutrition department, Hôtel-Dieu hospital, Paris, France. She is specialized in Endocrinology, Nutrition and Metabolic diseases. In 1996 she completed a PhD at Paris VII University related to the genetics of obesity in the French population. KC has oriented her carrier towards a specific profile as a bridge between clinical investigation and biological aspects of metabolic diseases. Her work led notably to the identification of monogenic forms of obesity and she contributed to more than 100 publications in international journals. During a postdoctoral stay at Stanford University, she acquired new competencies with the goal of approaching in a more integrative way the physiopathology of complex diseases using the gene profiling approach. In 2001 she obtained a young INSERM "Avenir" team focused on the characterization of patterns of gene expression induced by genetic or environmental perturbations that act on the energy balance equation. Since 2006 she's leading the INSERM U755 "Nutriomics" unit which focuses human obesity through genetic and transcriptomic approaches. She is a member of several national committees in obesity and metabolism (AFERO and ALFEDIAM).

Jean-Daniel Zucker is leading a Machine Learning team in the INSERM U755 "Nutriomics" unit which is mixed with Paris VI University. He is a former Aeronautical Engineer (Sup'Aéro, 1985) who specialized in a Master of Computing for Life Sciences (University Paris V, 1986). He worked in research and development for three years at the New England Medical Center (Boston, USA) and three years at Thomson-Syseca (St-Cloud, France). After a Master in

Artificial Intelligence in 1992, he got his PhD in 1996 in Machine Learning from Paris VI University where he became an associate professor focusing on representation changes and abstraction in learning. In 2002 he became Full Professor at Paris XIII University where he led a CNRS EPML team until the end of 2005. He is now co-director of the LIM&BIO (the Medical Informatics and Bioinformatics Laboratory of Paris XIII University). His research focus is on feature selection, feature construction and reformulated concept learning algorithm in Machine Learning for transcriptomics data.